The Road to Master Data Management: Identifying Master Data

December 13, 2006
David Loshin
Knowledge Integrity, Inc.
What is “Master Data”?  

- Core business objects used in the different applications across the organization, along with their associated metadata, attributes, definitions, roles, connections, and taxonomies, e.g.:  
  - Customers  
  - Suppliers  
  - Parts  
  - Products  
  - Locations  
  - Contact mechanisms  
  - ...
Transactions Use Master Data

“David Loshin purchased seat 15B on US Airways flight 238 from Baltimore (BWI) to San Francisco (SFO) on July 20, 2006.”

<table>
<thead>
<tr>
<th>Master Data Object</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer</td>
<td>David Loshin</td>
</tr>
<tr>
<td>Product</td>
<td>Seat 15B</td>
</tr>
<tr>
<td>Flight</td>
<td>238</td>
</tr>
<tr>
<td>Location</td>
<td>BWI</td>
</tr>
<tr>
<td>Location</td>
<td>SFO</td>
</tr>
</tbody>
</table>
MDM Objectives

An MDM program is intended to:

- Assess the use of core information objects, data value domains and business rules
- Identify core information objects relevant to business success
- Instantiate a shared standardized object model
- Manage collected and discovered metadata as an accessible, browsable resource
- Collect and harmonize replicated data instances
- Integrate the harmonized view via a service-oriented approach
- Institute the proper data governance policies and procedures at the corporate or organizational level
Structures and Semantics Vary Across the Enterprise

CUSTOMER

FirstName VARCHAR(14)
MiddleName VARCHAR(14)
LastName VARCHAR(30)
TelNum NUMERIC(10)

CUST

First VARCHAR(15)
Middle VARCHAR(15)
Last VARCHAR(40)
Address1 VARCHAR(45)
Address2 VARCHAR(45)
City VARCHAR(30)
State CHAR(2)
ZIP CHAR(9)
Finding Master Data

- What data elements constitute our “master data”?
- How do we locate and isolate master data objects that exist within the enterprise?
- How do we assess the variances between the different representations in order to consolidate instances into a single view?
Centralizing Semantics – Basic Questions

- Master data will be distributed and replicated across the application environments
- The initial goal is to consolidate the replicated copies into a single repository
- Before materializing a single master record for any entity, one must be able to:
  - Discover which data resources may contain entity information
  - Understand which attributes carry identifying information
  - Extract identifying information from the data resource
  - Transform the identifying information into a standardized or canonical form
  - Establish similarity to other standardized records
“A customer is an individual who has purchased one of our products”

“...individual to whom we have delivered one of our products”
Master Object Resolution

- Resolution of candidate master data types requires a compete view of what composes the information architecture
- This entails cataloging data sets, their attributes, data domains, definitions, contexts, and semantics
- This view must facilitate the resolution of:
  - *Format* at the element level,
  - *Structure* at the instance level, and
  - *Semantics* across all levels
- This introduces three challenges:
  - Collecting and analyzing master metadata
  - Resolving similarity in structure
  - Understanding and unifying master data semantics
Collecting Master Metadata

- Metadata sources:
  - Data dictionaries
  - E/R models
  - COBOL copybooks
  - Subject matter experts

- Data profiling – an effective approach to accumulating metadata

- Document all data element characteristics within a metadata repository using a standard representation
Analyzing Master Metadata

- The consolidated metadata repository will enumerate data characteristics in a standardized way.
- The standard representation enables us to subject the metadata to the same kinds of statistical analyses provided by data profiling tools, such as:
  - Assessing frequency of occurrence of names
  - Comparing the lengths of name fields
  - Discovering dependencies between attribute names and assigned types
Resolving Similarity in Structure

- Different underlying master meta-models are likely to share many commonalities
- Structures will probably reflect similar collections of attributes and relations
  - Example: Many customer data sets will contain customer names along with variant contact mechanisms such as addresses or telephone numbers
- The data analyst can review existing models to identify similar object structures
- Two “hints” in structural similarity:
  - Overlapping structures may reflect data sets carrying identifying information with variant sets of attributes
  - Derived structures potentially reflect an embedding of core master data attributes within a specialized version of the same object
Unifying Master Data Semantics

- What is the qualitative difference between pure syntactic/structural metadata and its underlying meaning?
- Review semantic consistency in data element naming and data types, sizes, and structures
- Next, document business meanings:
  - What are the data element definitions?
  - Are there authoritative sources for these definitions?
  - Do similar objects have different business meanings?
- Resolution of variance will help in standardizing both the representations and meanings for identified master data object types
Identifying and Qualifying Master Data Objects

- Collected semantic metadata simplifies the task
- Seek out similarity in syntax, structure, and meaning
- Reduce the complexity of analysis by simplifying classifications and loosening constraints
- For example, characterize each data element as one of
  - Flag
  - Time stamp
  - Magnitude
  - Code
  - Identifier
  - Cross-reference
Metadata Fractals

- Each data attribute can be summarized by a set of descriptive characteristics (e.g., data type, length, value class)
- All metadata is represented in a standardized format
- We are seeking similar data sets with similar structures, formats, and semantics – “identifying attribution”
- In reality, we are performing a search for objects that “match” or “duplicate” each other
An Interesting Realization

- Consider this:
  - We are using our tools to assess data element structure
  - We are collecting this information into a metadata repository
  - We use our tools to look for data attributes that share similar characteristics
  - We use our tools to seek out attributes with similar names
  - We analyze the data value sets and assign them into value classes
  - We use our tools to detect similarities between representative data meta-models

- In essence, the techniques and tools we can use for determining the sources of master data objects are the same types of tools we use for consolidating the data into a master repository!
Questions?

- If you have questions, comments, or suggestions, please contact me
  
  David Loshin  
  301-754-6350  
  loshin@knowledge-integrity.com